Vocabulary Scaling Law : Tuning Open-vocabulary Predictors for Their Openness (Appendix Included)

Ziliang Chen[1], Yulu Li[2], Liangda Fang[2], Jusheng Zhang[3]*, Yongsen Zheng[4], Quanlong Guan[2], Xipeng Chen[1]

[1]Research Institute of Multiple Agents and Embodied Intelligence, Peng Cheng Laboratory, [2]Jinan University, [3]Sun Yat-sen University, [4]NTU

Abstract

Open-vocabulary learning on CLIP provides remarkable generalization on diverse concepts, however, falters under the realistic streaming open-world evaluations for Stability against distractor classes and Extensibility to novel classes. Current fine-tuning methods often fail these tests since they are mainly designed for closed-set conditions, leading to the performance gaps while the target vocabulary progressively scales. We formalize a "vocabulary scaling law" showing that these openness measures can be lower-bounded by performance on the full class-name universe, implying that robust fine-tuning should: (i) account for the entire vocabulary, (ii) tune class-name embeddings rather than context, and (iii) enforce orthogonality between prompt embeddings including training and open-set class names. Guided by our analysis, we propose Submodular-Vocabulary Fine-tuning (SVFT), a bi-level optimization framework that approximates the intractable objective of tuning all class name embedding by greedily selecting a small, informative subset of class names via constrained submodular maximization, thus, allows the employment of efficient greedy algorithm for the near-optimal class-name subset selection to fine-tune CLIP instead of using all open classes. Across extensive experiments, SVFT consistently improves both stability and extensibility, advancing the openness and practical robustness of CLIP-based vision–language models.

1. Introduction

Vision-Language Models (VLMs) such as CLIP [9, 21] have marked a paradigm shift in visual recognition, leveraging natural language supervision from massive image-text datasets to enable incredible few-shot and even zero-shot inference [35, 36]. Their hallmark capability, open-vocabulary prediction, allows for the classification of images using arbitrary, user-defined category names, breaking free from the constraints of predefined, fixed-class datasets. This flexibility has positioned VLMs as a basic technology for a wide array of downstream applications that demand generalization to diverse and unforeseen visual concepts.

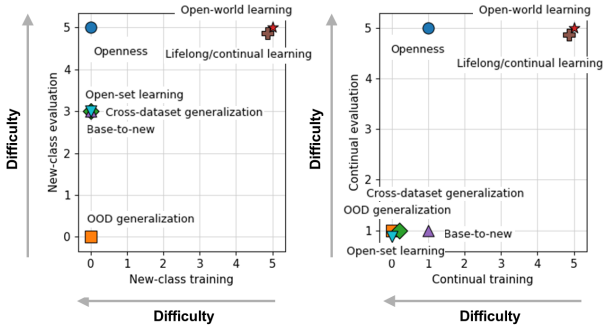

Figure 1. The comparison between fine-tuning for openness and the other task setups for CLIP from the aspects of whether there are new classes join the training and evaluation, as well as whether the training and evaluation are continually executed.

Despite their successes, a critical gap remains between their typical evaluation and the demands of real-world deployment. In particular, most fine-tuning schemes and evaluation protocols for CLIP operate under the unrealistic assumption that **the evaluated images are exactly consistent with the classes to construct the open vocabulary**. In other words, existing open-vocabulary predictors are mostly evaluated with close-set classes in the stationary setups. In the pursuit of the openness of CLIP (Fig.1), the concepts of **stability** and **extensibility** have emerged as more rigorous metrics for re-evaluating CLIP. Stability measures a model's robustness to maintain accuracy on known classes when the vocabulary is expanded with unseen "distractor" concepts, while extensibility measures its zero-shot ability to correctly classify close-set and open-set categories along with the same vocabulary expansion (Fig.2). As shown by [22], the CLIP family and their fine-tuning methods degrade significantly on these metrics while the vocabulary scaling, severely hampering their utility in an open world.

To demystify the reasons why they fail, our paper begins by investigating the underlying principles that govern CLIP's performance in an open-vocabulary context. We formalize a **"vocabulary scaling law"**, revealing that a

*indicate corresponding author;

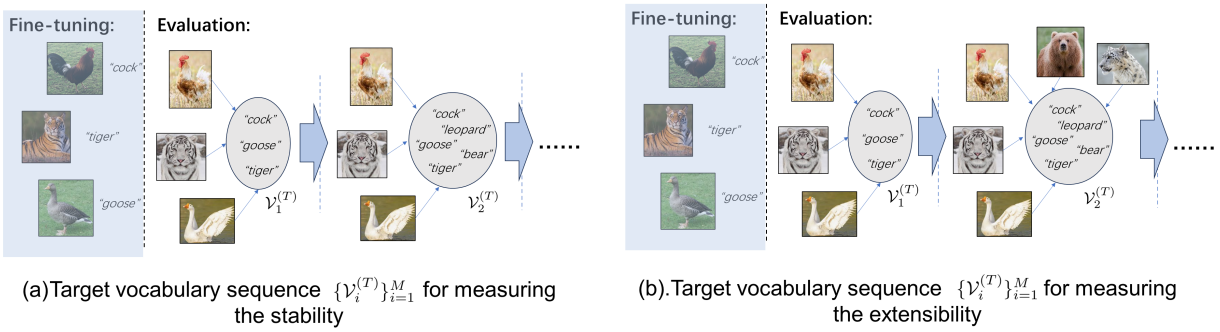

Figure 2. The illustration of stability metric (a) and extensibility metric (b) proposed to quantify the openness of CLIP. Specifically, the stability measures the model's ability to maintain performance on its known, fine-tuned classes when the evaluation vocabulary is expanded with novel, unseen "distractor" classes. In contrast, the extensibility measures the model's zero-shot generalization capability to correctly classify images of new classes that were not part of the fine-tuning set. Together, these metrics assess a model's robustness in an open-world setting: it should not forget what it knows (stability) while being able to adapt to what it has never seen (extensibility).

model's prediction confidence for a correct class is inherently lower-bounded by its performance on the scaled vocabulary that includes all possible open-set classes. This leads to **Takeaway 1**: to maximize stability and extensibility on known concepts, ***fine-tuning should not be limited to the target training vocabulary*** $\mathcal{V}^{(T)}$ ***but must account for the entire open-vocabulary universe*** U. However, this presents a critical trade-off, as naively adjusting the model for all classes in U can corrupt the carefully learned embeddings for unseen classes, thereby harming zero-shot generalization. Our analysis also reveals (**Takeaway 2**) that it can be mitigated by ***exclusively fine-tuning the class-name embeddings for*** $\mathcal{V}^{(T)}$ ***while enforcing an orthogonality constraint against the embeddings of classes in*** $U/\mathcal{V}^{(T)}$**.**

Enlighten by these insights, we propose a novel approach Submodular-Vocabulary Fine-tuning (SVFT) to make this theoretically-grounded optimization practical. In particular, fine-tuning CLIP on the entire class universe is computationally infeasible due to its immense size. SVFT overcomes this through a continuous-discrete bi-level optimization framework. At each step, instead of using all open-vocabulary classes, our approach selects a small yet highly representative subset of classes for effectively approximating the full objective on the vocabulary universe. We demonstrate that this class-name subset selection can be formally cast as a submodular function maximization problem under the cardinality contraint. This key connection allows us to employ an efficient greedy-search algorithm to find a near-optimal class-name subset conditioned on the prompt-based feedback with regards to their loss inferences, making our approach both scalable and highly effective.

Our contributions are threefold: **(1)**. we provide the first formal analysis of the vocabulary scaling law, the key rules to achieve the openness of CLIP; **(2)**. we develop the SVFT framework, which operationalizes these insights into a practical and efficient class-name fine-tuning strategy; **(3)**. we establish the connection between open-vocabulary learning and submodularity. In extensive experiments, we verify the vocabulary scaling laws from diverse aspects, then demonstrate that the extraordinary performance of SVFT to the robustness on both stability and extensibility.

2. Related Work

CLIP and open-vocabulary prediction. CLIP and its variants [9, 21, 25, 26, 37], emerge as a breakthrough in transferring visual representations through natural language supervision, enabling generalization across diverse visual recognition tasks [2, 7, 10, 32]. It establishes contrastive pre-training on massive image-text pair datasets to facilitate open-vocabulary prediction, *i.e.*, a prompt template such as "a photo of a [CLASS]," with any potential category name can be semantically encoded as the category-specific classification weight. A parallel research direction [1, 17, 28] explores sequence-to-sequence generation rather than contrastive discrimination for open-vocabulary recognition, beyond our scope.

Open-set and open-world learning. Open-set learning [5, 13, 23, 27, 29, 30, 38, 39] challenges classification systems to identify samples from categories not seen during training as "unknown." It is closely related either as one-versus-rest classification [23] or as multi-class classification [14, 24] tasks. Open-world learning [3, 20] extends it by additionally requiring systems to incrementally incorporate newly labeled unknown instances into an expanding classification framework. Distinct from these paradigms, CLIP-driven open-vocabulary prediction can operate inference in a post-training-free manner. However, [22] verified that if open-vocabulary prediction incorporate more open-set class names without any target dataset training, the zero-shot in-

ference performance by CLIP will gradually drops in terms of the extensibility and the scalability metrics. It significantly hampers the validity of CLIP since we can not expect to a lexical list before classifying objects in the open world.

Open-vocabulary fine-tuning. Instead of zero-shot inference, partially fine-tuning CLIP with a few of target training data can yield more powerful open-vocabulary prediction models. The fine-tuning techniques can be divided into three research lines: (1) Adapter-based tuning [12, 31, 34]: inserting an extra tiny layer/module to the frozen vision or text encoders, which are fine-tuned in terms of the target training set; (2) Prompt-tuning [6, 35, 36]: fine-tuning the context embedding parameters of the prompt template to classifying the target images; (3) Name-tuning [18, 19]: directly fine-tuning the category-specific parameter if we need to recognize the images with respect to this category. Some approaches [15, 18] fall into their intersection to reap their both advantages.

3. Background: the Openness of CLIP

In this section, we provide the preliminaries of CLIP model and how to evaluate its openness by *stability* and *extensibility*, the critical metrics to facilitate our follow-up analysis.

CLIP-based open-vocabulary prediction. Suppose f,g be the vision and text encoders well-trained by CLIP. The model accepts an image $\mathbf{x}$ along with a set of candidate categories whose names are summarized into a *target vocabulary* $\mathcal{V}^{(T)} = \{w_i\}_i^{\mathbf{m}}$ (w_i denotes the i-th class name in $\mathcal{V}^{(T)}$). Then the predictor $P_{f,g}$ classifies the image $\mathbf{x}$ by

$$\begin{aligned}\hat{y} &= \underset{i\in\mathcal{V}^{(T)}}{\operatorname{argmax}}\, P_{f,g}\big(\mathbf{x}, \mathcal{V}^{(T)}\big) \\ &= \underset{i\in\mathcal{V}^{(T)}}{\operatorname{argmax}}\, \frac{\exp\left(\langle g(\boldsymbol{T}(e(w_i))), f(\mathbf{x})\rangle/\gamma\right)}{\sum_{k=1}^{\mathbf{m}} \exp\left(\langle g(\boldsymbol{T}(e(w_k))), f(\mathbf{x})\rangle/\gamma\right)},\end{aligned} \tag{1}$$

where $\boldsymbol{T}(e(w_i))$ denotes the textual prompt embedded from "a photo of a w_i", where $e(w_i)$ indicates the class-name embedding of the class name w_i and $\boldsymbol{T}(\cdot)$ represents the embedding of context format, therefore we isolate the post-training processes of $e(w_i)$ and $\boldsymbol{T}$ into *class-name* prompt-tuning and *context-based* prompt-tuning, respectively. $\langle\cdot,\cdot\rangle$ denotes cosine similarity the between the normalized image feature $f(\mathbf{x})$ and the normalized prompt feature $g(\boldsymbol{T}(e(w_i))$ produces open-vocabulary prediction $P_{f,g}$ that uses $g(\boldsymbol{T}(e(w_i))$ as the classifier's weight vector with respect to the class w_i.

Vocabulary stability and extensibility. The evaluation of $P_{f,g}$'s few-shot and zero-shot inference are almost conducted under the assumption that the test-set image $\mathbf{x}$ belong the classes in the target vocabulary (Fig.2 (a-b)). The unpractical setup motivates the exploration to CLIP's openness to the vocabulary including more other categories that all test images do not belong to, which refers to the stability and the extensibility metrics derived from [22]. Specifically, suppose that $\boldsymbol{D}_{\mathcal{V}}$ denotes the dataset labeled by the classes in a vocabulary $\mathcal{V}$, and $\{V_i\}_{i=1}^{M}$ represents a series of augmenting vocabularies that each holds the consistent size $\mathbf{m}$ with the original target vocabulary $\mathcal{V}^{(T)}$ (*i.e.*, $\forall i\in[1{:}M]$, $|V_i|{=}\mathbf{m}$). These vocabularies satisfy that $V_1 = \mathcal{V}^T$ and $\forall i, j\in[1{:}M]$, $V_i \cap V_j = \emptyset$ if $j \neq i$ (*i.e.*, no training images are categorized into the classes in $\{V_i\}_{i=2}^{M}$). Then we can define the series of scaling (target) vocabularies $\{\mathcal{V}_i^{(T)}\}_{i=1}^{M}$ through the ***vocabulary scaling process***

$$\mathcal{V}_i^{(T)} = \cup_{j=1}^{i} V_j,\ \text{s.t.}\ V_j \sim \boldsymbol{P}_V(\cdot|\mathcal{V}_{i-1}^{(T)}),\ \forall i \in [M]. \tag{2}$$

where $\boldsymbol{P}_V(\cdot|\mathcal{V}_{i-1}^{(T)})$ indicates the discrete stochastic process that draw $\mathbf{m}$ classes beyond the previous-stage target vocabulary $\mathcal{V}_{i-1}^{(T)}$, and $\boldsymbol{P}_V(V_1|\mathcal{V}_0^{(T)}) = \mathbf{1}_{V_1=\mathcal{V}^{(T)}}$. Given this, the ***stability*** $\mathrm{ACC}_{\mathrm{S}}(\mathcal{V}^{(T)})$ and ***extensibility*** $\mathrm{ACC}_{\mathrm{E}}(\mathcal{V}^{(T)})$ can be computed by drawing multiple vocabulary sequences from $\boldsymbol{P}_V(\cdot|\mathcal{V}_{i-1}^{(T)})$, then average their mean accuracy at each step i, *i.e.*,

$$\begin{aligned}&\mathrm{ACC}_{\mathrm{S}}(\mathcal{V}^{(T)}) := \mathbb{E}_{\{V_j\}_{j=1}^{M}\sim\boldsymbol{P}_V, \mathcal{V}_i^{(T)}=\cup_{j=1}^{i}V_j^{(T)}} \\ &\frac{1}{M}\sum_{i=1}^{M}\Big(\mathbb{E}_{<\mathbf{x},\mathbf{y}>\sim\boldsymbol{D}_{\mathcal{V}^{(T)}}}\mathbb{I}\Big(\mathbf{y} = \underset{j\in\mathcal{V}_i^T}{\operatorname{argmax}}\, P_{f,g}\big(\mathbf{x}, \mathcal{V}_i^{(T)}\big)\Big)\Big);\end{aligned} \tag{3}$$

$$\begin{aligned}&\mathrm{ACC}_{\mathrm{E}}(\mathcal{V}^{(T)}) := \mathbb{E}_{\{V_j\}_{j=1}^{M}\sim\boldsymbol{P}_V, \mathcal{V}_i^{(T)}=\cup_{j=1}^{i}V_j^{(T)}} \\ &\frac{1}{M}\sum_{i=1}^{M}\Big(\mathbb{E}_{<\mathbf{x},\mathbf{y}>\sim\boldsymbol{D}_{\mathcal{V}_i^{(T)}}}\mathbb{I}\Big(\mathbf{y} = \underset{j\in\mathcal{V}_i^{(T)}}{\operatorname{argmax}}\, P_{f,g}\big(\mathbf{x}, \mathcal{V}_i^{(T)}\big)\Big)\Big).\end{aligned} \tag{4}$$

where $\forall\{V_j\}_{j=1}^{M} \sim \boldsymbol{P}_V$, the stochastic process holds $U = \cup_{j=1}^{M} V_j$ with the vocabulary universe U to include all open classes for the classification interference. It is noteworthy that, the original formulations of stability and extensibility are reframed into their equivalent presentation in Eq.3,4, in order to facilitate our analysis. We certify their consistency in our Appendix.B.

4. Vocabulary Scaling Law behind CLIP

The concepts of stability and extensibility significantly challenge the CLIP's performances upon the in-distribution (ID) and out-of-distribution (OOD) generalization evaluated in the conventional manner. Concretely, they assesses the evaluation by progressively introducing open-set classes to perturb the prompt-based classification accuracy. While such measures are crucial to capture CLIP's performance in an open-world context, existing CLIP-based open-vocabulary predictors have notable shortcomings: they either rely exclusively on the zero-shot inference endowed by pre-training, or they undergo few-shot fine-tuning without explicit consideration for stability and extensibility.

In this section, we discussed how the stability and extensibility of CLIP change along with the target vocabulary $\mathcal{V}^{(T)}$ scaling demonstrated in Eq.2. It leads to some takeaways to customize existing CLIP-based few-shot fine-tuning approaches to improve the stability and the extensibility. They enlighten our methodology in the next section.

A simple lower bound of openness. Let's consider each augmenting open-vocabulary sequence $\{V_j^{(T)}\}_{j=1}^{M}$ drawn from the stochastic process $\boldsymbol{P}_V$. Based on formulations 3,4, there are some significant findings observed:

- The sequence of open-vocabulary prediction $\{P_{f,g}\big(\mathbf{x}, \mathcal{V}_i^{(T)}\big)\}_{i=1}^{M}$ determines $\mathrm{ACC_S}$ with regards to $\boldsymbol{P}_V$.
- $\mathrm{ACC_E}$ holds a equivalent decomposition as

$$\mathrm{ACC_E}(\mathcal{V}^{(T)}) := \mathbb{E}_{\{V_j\}_{j=1}^{M}\sim \boldsymbol{P}_V, \mathcal{V}_i^{(T)}=\cup_{j=1}^{i} V_j^{(T)}} \Bigg(\frac{1}{M}\sum_{i=1}^{M}\Big(\mathbb{E}_{<\mathbf{x},\mathbf{y}>\sim \boldsymbol{D}_{\mathcal{V}^{(T)}}} \mathbb{I}\Big(\mathbf{y} = \underset{j\in\mathcal{V}_i^{(T)}}{\operatorname{argmax}} P_{f,g}\big(\mathbf{x}, \mathcal{V}_i^{(T)}\big)\Big)\Big) + \frac{1}{M-1}\sum_{i=2}^{M}\Big(\mathbb{E}_{<\mathbf{x},\mathbf{y}>\sim \boldsymbol{D}_{\cup_{j=1}^{i} V_j}} \mathbb{I}\Big(\mathbf{y}= \underset{k\in\cup_{j=1}^{i} V_j}{\operatorname{argmax}} P_{f,g}\big(\mathbf{x}, \mathcal{V}_i^{(T)}\big)\Big)\Big)\Bigg) \tag{5}$$

where the first term exactly refers to the stability $\mathrm{ACC_S}$. Eq.5 illustrates the optimization focuses on stability and extensibility could be coupled on the classes in $\mathcal{V}^{(T)}$. To this, fine-tuning with few-shot instances drawn from $\boldsymbol{D}_{\mathcal{V}^{(T)}}$ can potentially outperform CLIP's zero-shot inference on $\mathrm{ACC_S}$ and $\mathrm{ACC_E}$.

Provided the observations, for each $<\mathbf{x},\mathbf{y}>\sim \boldsymbol{D}_{\mathcal{V}^{(T)}}$, we derive an inequality chain on $\{P_{f,g}(\mathbf{x}, \mathcal{V}_i^{(T)})\}_{i=1}^{M}$:

Proposition 1. *$\forall <\mathbf{x},\mathbf{y}> \in \boldsymbol{D}_{\mathcal{V}^{(T)}}$, and $w_{\mathbf{y}}$ indicates the class name of $\mathbf{y}$ that $\mathbf{x}$ belongs to. $\forall\{V_i\}_{i=1}^{M} \sim \boldsymbol{P}_V$, it holds*

$$P_{f,g}^{(w_{\mathbf{y}})}(\mathbf{x}, \mathcal{V}_M^{(T)}) \leq \cdots \leq P_{f,g}^{(w_{\mathbf{y}})}(\mathbf{x}, \mathcal{V}_2^{(T)}) \leq P_{f,g}^{(w_{\mathbf{y}})}(\mathbf{x}, \mathcal{V}_1^{(T)}) \tag{6}$$

where

$$P_{f,g}^{(w_{\mathbf{y}})}(\mathbf{x}, \mathcal{V}_i^{(T)}) = \frac{\exp\left(\langle g(\boldsymbol{T}(e(w_{\mathbf{y}}))), f(\mathbf{x})\rangle/\gamma\right)}{\sum_{k=1}^{|\mathcal{V}_i^{(T)}|}\exp\left(\langle g(\boldsymbol{T}(e(w_k))), f(\mathbf{x})\rangle/\gamma\right)} \tag{7}$$

Proposition 1 implies a critical improved technique to optimize CLIP for the stability and extensibility. First, minimizing $\mathbb{E}_{<\mathbf{x},\mathbf{y}>\sim \boldsymbol{D}^{\mathsf{train}}_{\mathcal{V}^{(T)}}} - \log P_{f,g}^{(w_{\mathbf{y}})}(\mathbf{x}, \mathcal{V}_M^{(T)})$ sufficiently improve all the classification results across the scaling vocabularies in $\{\mathcal{V}_i^{(T)}\}_{i=1}^{M}$ because $P_{f,g}^{(w_{\mathbf{y}})}(\mathbf{x}, \mathcal{V}_M^{(T)})$ typically determine the lower bound of their vocabulary prediction in Eq.5. Second, despite the scaling vocabulary sequence $\{\mathcal{V}_i^{\pi}\}_{i=1}^{M}$ varying with regards to the uncertain order of open-class names drawn from the stochastic process $\boldsymbol{P}_V^{(T)}$, $\mathcal{V}_M^{(T)}=\cup_{j=1}^{M} V_j^{\pi}=U$ are completely deterministic on the vocabulary universe U. Combining the result in Eq.5, we know that $\mathrm{ACC_S}(\mathcal{V}^{(T)})$ and $\mathrm{ACC_E}(\mathcal{V}^{(T)})$ on $\boldsymbol{D}_{\mathcal{V}^{(T)}}$ are both lower bounded by $P_{f,g}^{(w_{\mathbf{y}})}(\mathbf{x}, U)$.

> **Takeaway 1.** Training $P_{f,g}^{(w_{\mathbf{y}})}(\mathbf{x}, U)$ to classify $\mathbf{x}\sim \boldsymbol{D}_{\mathcal{V}^{(T)}}$ improves the stability, and also improves the part of the extensibility with respect to $\mathcal{V}^{(T)}$.

The takeaway explains the sub-optimal of existing CLIP-based approaches on the openness metrics, since their fine-tuning are limited in classifying $\mathcal{V}^{(T)}$ yet never incorporate the names of the open classes in $U/\mathcal{V}^{(T)}$. Proposition.1 implies that their performances on the stability and extensibility consistently decrease along with more open classes progressively included by scaling the target vocabulary.

Why tuning class-name embedding matters for extensibility? It is noteworthy that in the open-world mechanism, training images with their class names in $\{V_i\}_{i=2}^{M}$ are unavailable, thus, CLIP's performance on the extensibility $\mathrm{ACC_E}(\mathcal{V}^{(T)})$ with respect to $\{\boldsymbol{D}_{\cup_{j=1}^{i} V_j}\}_{i=2}^{M}$, (*i.e.*, the second term in Eq.5) depends upon the zero-shot inference ability obtained in the pre-training stage. With this regards, we elaborate a test instance in the second term in Eq.5

$$\underset{w(\mathbf{y})\in\cup_{j=2}^{i} V_j}{\operatorname{argmax}} P_{f,g}\big(\mathbf{x}, \mathcal{V}_i^{(T)}\big) \text{ s.t. } \forall i\in[2:M], \forall\langle\mathbf{x},\mathbf{y}\rangle\sim \boldsymbol{D}_{\cup_{j=1}^{i} V_j} = \underset{w(\mathbf{y})\in\cup_{j=2}^{i} V_j}{\operatorname{argmax}} \frac{\exp\left(\frac{\langle g(\mathbf{p}_{\mathbf{y}}), f(\mathbf{x})\rangle}{\gamma}\right)}{\sum_{\substack{w(\mathbf{y}')\\ \in\cup_{j=2}^{i} V_j}}\exp\left(\frac{\langle g(\mathbf{p}_{\mathbf{y}}), f(\mathbf{x})\rangle}{\gamma}\right) + \sum_{\substack{w(\mathbf{y}')\\ \in\mathcal{V}^{(T)}}}\exp\left(\frac{\langle g(\mathbf{p}_{\mathbf{y}'}), f(\mathbf{x})\rangle}{\gamma}\right)}, \tag{8}$$

where $\mathbf{p}_{\mathbf{y}}=\boldsymbol{T}(e(w(\mathbf{y})))$. From the elaboration, we realize that no matter of fine-tuning the CLIP model or the prompt-context embedding for $P_{f,g}^{(w_{\mathbf{y}})}(\mathbf{x}, U)$ to pursuit our first takeaway, the first term and the second term in the denominator of Eq.8 (the derivation of $P_{f,g}\big(\mathbf{x}, \mathcal{V}^{(T)}\big)$) would simultaneously change along with increasing the first term in the elaborated extensibility in Eq.5, so that only Takeaway 1 can not guarantee the overall improvement of $\mathrm{ACC_E}$.

Instead, if we only tune the $\mathcal{V}^{(T)}$-specific class-name embedding using $\boldsymbol{D}^{(\mathsf{train})}_{\mathcal{V}^{(T)}}$, *i.e.*, $\{e(w(\mathbf{y}))|\forall w(\mathbf{y})\in\mathcal{V}^{(T)}\}$, only the second term in the denominator of Eq.8 will change. In this case, decreasing $\exp\left(\frac{\langle g(\mathbf{p}_{\mathbf{y}'}), f(\mathbf{x})\rangle}{\gamma}\right)$, $\forall\mathbf{y}'\in\mathcal{V}^{(T)}$; $\forall\mathbf{x}\in\boldsymbol{D}_{\cup_{j=1}^{i} V_j}$, can encourage the update consistently improving the second term in Eq.5. Since $\boldsymbol{D}^{\mathsf{train}}_{\cup_{j=1}^{i} V_j}$ are unavailable, we turn to encourage the decrease between the class embedding sets in $\mathcal{V}^{(T)}$ and $U/\mathcal{V}^{(T)}$.

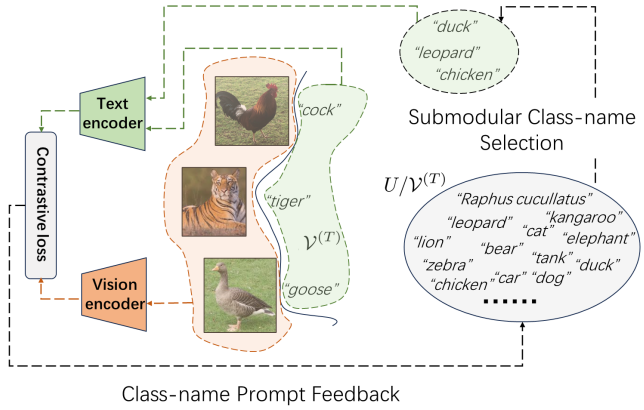

Figure 3. The pipeline of VSFT. It is formulated into a bi-level optimization where it first performs submodular class-name selection to greedily choose a small, informative subset of class names from a large, open-vocabulary universe. This selected subset joins the class-name prompt-loss feedback loop to fine-tune the class-name embedding. The process allows efficiently learning CLIP from a vast number of concepts to improve its stability and extensibility.

Takeaway 2. The class-name prompt tuning aims to classify $< \mathbf{x}, \mathbf{y} > \sim \boldsymbol{D}_{\mathcal{V}^{(T)}}$ into the class names in U with the orthogonal constraint between the class-query embedding in $\mathcal{V}^{(T)}$ and $U/\mathcal{V}^{(T)}$, encouraging the extensibility improvement.

5. Submodular-Vocabulary Fine-tuning

In the previous section, we elaborate the factors behind CLIP to steer its model robustness towards the vocabulary scaling process $\boldsymbol{P}_V$. As demonstrated by Takeaway 1, 2, we further formulate the class-name prompt-tuning objective to achieve their goals:

$$\min_{\{e(w)\}_{\forall w\in U}} \mathbb{E}_{\substack{<\mathbf{x},\mathbf{y}> \\ \sim \boldsymbol{D}^{\text{train}}_{\mathcal{V}^{(T)}}}} \Big[-\log P^{(w_{\mathbf{y}})}_{f,g}(\mathbf{x}, U) + \lambda \mathbb{E}_{\substack{w\in\mathcal{V}^{(T)}, \\ w'\in U/\mathcal{V}^{(T)}}} \langle g\big(\boldsymbol{T}(e(w))\big), g\big(\boldsymbol{T}(e(w'))\big)\rangle \Big] \tag{9}$$

where $\{e(w)\}_{\forall w\in U}$ denotes the set including all the class-name embedding drawn from the open-vocabulary universe U. Note that $\{e(w)\}_{\forall w\in U}$ has to be emitted with their contexts so that we prefer tuning their query prompt embedding $g\big(\boldsymbol{T}(e(w))$ with frozen context embedding; the constraint in the second term demonstrates that the optimization of class-name embedding in $\mathcal{V}^{(T)}$ (the first term) should not go to the direction with the effect on any open classes in $U/\mathcal{V}^{(T)}$, and λ denotes the hyper-parameter to control the trade-off between stability and extensibility.

The customized open-vocabulary learning in Eq.9, however, could be infeasible in practice: the optimization requires fine-tuning the embedding vectors of all the classes in U. It implies the large-scale training with a huge number of embedding because U should contains a sufficiently large number of classes to promise the openness evaluation. To this end, we prefer using a class subset $S \subset U$ instead of all the classes in U, therefore re-formulate Eq.9 into a new bi-level optimization objective, thus,

$$\begin{aligned} &\min_{\{e(w)\}_{\forall w\in S\cup\mathcal{V}^{(T)}}} \max_{S\subset U/\mathcal{V}^{(T)}, |S|\le K} F(\{e(w)\}, S) \\ &:= \mathbb{E}_{\substack{<\mathbf{x},\mathbf{y}> \\ \sim \boldsymbol{D}^{\text{train}}_{\mathcal{V}^{(T)}}}} \Big[-\log P^{(w_{\mathbf{y}})}_{f,g}(\mathbf{x}, S\cup\mathcal{V}^{(T)}) \\ &\quad + \lambda \mathbb{E}_{\substack{w\in\mathcal{V}^{(T)}, \\ w'\in S}} \Big(\langle g\big(\boldsymbol{T}(e(w))\big), g\big(\boldsymbol{T}(e(w'))\big)\rangle + 1 \Big) \Big] \end{aligned} \tag{10}$$

where the extra maximization is proposed to select K open classes from U (at most), such that the objective functional $F(\cdot, S)$ derived from each altering step could be representative enough to approximate the original objective functional $F(\cdot, U)$ with regards to the variation of class-name embedding through the optimization process. Besides, it is observed that the value of $\langle g\big(\boldsymbol{T}(e(w))\big), g\big(\boldsymbol{T}(e(w'))\big)\rangle$ range in $[-1, 1]$. Given this, the second term is not promised to be monotonic as the size of S increases, where the submodular maximization (we elaborate in the next section) can only be solved via the bi-level search [4] with the inferior approximation ratio. To this, we reconfigure each multiplied pair by $\Big(\langle g\big(\boldsymbol{T}(e(w))\big), g\big(\boldsymbol{T}(e(w'))\big)\rangle + 1\Big) \in [0, 2]$.Hence Eq.9 turns into a monotonic constrained submodular maximization objective (Eq.10) so that we could apply the more simple yet effective linear greedy search [16] to obtain the optimal S^* for each turn.

Optimization. When $\overline{S}$ has been determined, the objective reduced to $\min_{\{e(w)\}_{\forall w\in\overline{S}\cup\mathcal{V}^{(T)}}} F(\{e(w)\}, \overline{S})$ can be optimized by tuning the class-name embedding in $\overline{S}\cup\mathcal{V}^{(T)}$. Existing studies [8, 18, 19] have explored this prompt-tuning routine, which we adopt [8] into our implementation.

With the prompt-loss feedback updated in the previous iter, the bi-level optimization objective turns to select the optimal subset to achieve $\max_{S\subset U/\mathcal{V}^{(T)}, |S|\le K} F(\{\overline{e}(w)\}, S)$. It is non-trivial because the discrete optimization problem could not be solved by gradient descent, and there are no generic algorithms to solve the subset selection problem in the polynomial time. While interestingly, the class-set function $F(\{\overline{e}(w)\}, S)$ can be proved to satisfy the diminishing return, *i.e.*, also known as *submodularity* in the utility theory [11, 16]:

Definition 2. ***(Submodular function)*** A function $f : 2^V \to \mathbb{R}$ is *submodular* if for every $A \subseteq B \subseteq V$ and $e \in V \setminus B$ it holds that

$$\Delta_f(e \,|\, A) \ge \Delta_f(e \,|\, B).$$

where $\Delta_f(e \,|\, S) = f(S\cup\{e\}) - f(S)$ indicates the diminishing return property. Equivalently, a function $f : 2^V \to \mathbb{R}$

Table 1. The neural / vocabulary scaling laws evaluated by the comprehensive comparisons of different models on CIFAR100, ImageNet (Entity13), and ImageNet (Living17) datasets. The models are evaluated on closed-set accuracy (Acc-C), Extensibility (Acc-E), and Stability (Acc-S). The Δ columns show the performance drop relative to the closed-set accuracy.

Model	CIFAR100					ImageNet (Entity13)					ImageNet (Living17)				
	Acc-C	Extensibility		Stability		**Acc-C**	Extensibility		Stability		**Acc-C**	Extensibility		Stability	
		Acc-E	Δ	Acc-S	Δ		Acc-E	Δ	Acc-S	Δ		Acc-E	Δ	Acc-S	Δ
Neural Scaling Law:——							********								
CLIP (RN101)	68.3	55.4	-12.9	54.9	-13.4	80.4	77.4	-3.0	77.3	-3.1	77.6	74.5	-3.1	74.4	-3.2
CLIP (ViT-B/32)	78.0	69.6	-8.4	68.9	-9.1	80.8	78.0	-2.8	77.8	-3.0	78.0	74.4	-3.6	75.0	-3.0
CLIP (ViT-B/16)	**79.7**	**72.6**	**-7.1**	**72.0**	**-7.7**	**83.5**	**81.1**	**-2.4**	**81.0**	**-2.5**	**79.5**	**77.9**	**-1.6**	**77.6**	**-1.9**
SLIP (ViT-B/16)	63.9	51.1	-12.8	50.4	-13.5	65.7	62.3	-3.4	62.0	-3.7	65.7	62.6	-3.1	62.5	-3.2
DeCLIP (ViT-B/32)	**78.7**	**70.8**	**-7.9**	**70.4**	**-8.3**	**81.9**	**79.2**	**-2.7**	**79.1**	**-2.8**	**82.1**	**80.2**	**-1.9**	**80.0**	**-2.1**
PE (ViT-B/32)	78.3	70.3	-8.0	69.9	-8.4	81.9	79.4	-2.5	79.2	-2.7	78.7	76.0	-2.7	75.8	-2.9
PE (ViT-B/16)	**79.6**	**72.6**	**-7.0**	**72.0**	**-7.6**	**85.3**	**83.2**	**-2.1**	**83.1**	**-2.2**	**79.6**	**78.2**	**-1.4**	**78.0**	**-1.6**
Vocabulary Scaling Law(ViT-B/16):							********								
Context-based PT ($\mathcal{V}^{(T)}$)	83.6	76.9	-6.7	76.7	-6.9	87.5	85.3	-2.2	85.5	-2.0	82.7	82.6	-0.1	81.3	-1.4
Context-based PT (U)	84.1	73.2	-10.9	80.4	-2.7	89.2	86.6	-2.6	84.1	-2.5	83.2	82.6	-0.6	82.4	-0.9
Name-based PT ($\mathcal{V}^{(T)}$)	84.2	79.4	-4.8	77.8	-6.4	86.8	85.2	-1.6	85.9	**-0.9**	83.5	83.0	-0.5	81.8	-0.8
Name-based PT (U)	85.6	81.4	-4.2	82.8	-2.8	88.8	87.1	**-1.7**	**87.5**	-1.3	84.1	83.8	-0.3	83.5	**-0.6**
Name-based PT + Orth (U)	**87.8**	**83.7**	**-4.1**	**85.2**	**-2.6**	**90.2**	**88.3**	-1.9	87.3	-2.9	**86.2**	**86.1**	**-0.1**	**84.3**	-0.9

is also *submodular* if for every $A, B \subseteq V$,

$$f(A \cap B) + f(A \cup B) \leq f(A) + f(B).$$

The submodularity delivers many useful technical characteristic to set functions, particularly, its maximization under the cardinality constraint can provide the ϵ-approximate result to the global optima. Combine them and we have

Theorem 3. ***(Submodular Class-name Selection)*** *Suppose that $F(\{\overline{e}(w)\}, S)$ denotes the class-name set function with respect to $\forall S \subset U/\mathcal{V}^{(T)}$. Then $F(\{\overline{e}(w)\}, S)$ is submodular* w.r.t. *the universe $U/\mathcal{V}^{(T)}$, therefore $\max_{S \subset U/\mathcal{V}^{(T)}, |S| \leq K} F(\{\overline{e}(w)\}, S)$ refers to a submodular maximization problems with the K-size cardinality constraint, and its global optimal subset S^* can be approximated with $\hat{S}$ obtained by a greedy-search algorithm:*

$$F(\{\overline{e}(w)\}, \hat{S}) \geq (1 - \frac{1}{e}) F(\{\overline{e}(w)\}, S^*). \tag{11}$$

Hence our customized open-vocabulary learning strategy is named by *Submodular-vocabulary Fining-tuning* (SVFT). Fig.3 briefly illustrate the SVFT pipeline, and its implementation details are proposed in Appendix.B.

6. Experiments

In this section, we conducted comprehensive experiments to validate our theoretical justification in Sec.4 and the superiority of our SVFT approach (Sec.5).

6.1. Verification of Vocabulary Scaling Law

Understanding the CLIP-based classifier with the target vocabulary scaling via the sequence $\{\mathcal{V}_i^{(T)}\}_{i=1}^M$, refer to Takeaway 1,2 that jointly certify the mitigation of performance gaps in the stability and extensibility metrics. Despite their theoretical derivation, Takeaway 1,2 have never been verified in the empirical study. It motivates us to testify these principles. Their verification follows the setup in [22] including CIFAR100, ImageNet (Entity13), ImageNet (Living17) as the benchmarks, and the vocabulary construction details are found in our Appendix.B.

Baselines. We employ 5 baselines derived from Takeaway 1,2 to yield the thorough understanding: Context-based PT (prompt-tuning) with the original target vocabulary $\mathcal{V}^{(T)}$ [36]; Context-based PT with the vocabulary universe U to replace $\mathcal{V}^{(T)}$ in Eq.1; As a counterpart of Context-based PT, we also account for the baselines Class-name PT ($\mathcal{V}^{(T)}$) and Class-name PT (U) derived from the learning-to-name mechanism [8], where the prompt optimization focus on the relevant class-name embedding in $\mathcal{V}^{(T)}$ and U while the context embedding parameters in the class-query prompt $\boldsymbol{T}(\cdot)$ are frozen; then the last baseline Class-name PT + Orth (U) is derived from the Class-name PT (U) with the additional constraint to enforce the orthogonal relations of the class-query embedding between $\mathcal{V}^{(T)}$ and $U/\mathcal{V}^{(T)}$. It is noteworthy that SVFT involves the algorithm design beyond the discussion to vocabulary scaling law, hence we present its experiment later.

Ablation. The vocabulary scaling law concluded by Takeaway 1,2, can be generally interpreted via the three comparison cases across the baselines. (1) *Context-based / Class-name PT ($\mathcal{V}^{(T)}$)* v.s. *Context-based / Class-name PT (U)*; (2) Context-based PT (U) *v.s.* Class-name PT (U); (3) Class-name PT (U) *v.s.* Class-name PT + Orth (U). We report their original accuracy on $\mathcal{V}^{(T)}$, *i.e.* ACC the accuracy with regards to the stability $\mathrm{ACC_S}$ and extensibility $\mathrm{ACC_E}$, and their performance gaps $\triangle$ compared with the original accuracy, *i.e.*, $\triangle$=$\mathrm{ACC_S}$-ACC or $\triangle$=$\mathrm{ACC_E}$-ACC. As observed in Table.1, the results in the case (1) consistently

demonstrate the superiority of learning the class name embedding parameters compared with the context optimization strategies in Acc-C. It also yields better extensibility, in particular, when the target vocabulary is U, the extensibility performance gaps refer to -10.9 (Context-based) and -4.2 (Name-based); while the stability performance gaps are relatively subtle. The observations are typically aligned with the statement in Takeaway.1. In the case (2), we have broadly observe that the extensibility and the stability of Class-name PT (U) simultaneously exceed those presented by Context-based PT (U), while the comparison between Class-name PT and Context-based PT does not hold the significant evidences on the vocabulary $\mathcal{V}^{(T)}$. Note that the suggestion in Takeaway.2 is built on the fulfillment with the vocabulary prediction on U, henceforth the results are also consistent with Takeaway.2. Finally, the case (3) model almost outperform the other baselines, which sufficiently verifies the Takeaway 2's demonstration.

Neural scaling law *v.s.* vocabulary scaling law. Beyond the aforementioned ablation pairs in vocabulary scaling law, Table.1 we also intentionally reserve the zero-shot inference results of diverse CLIP-family models derived from [22], in particular, *e.g.*, CLIP, SLIP, DeCLIP, and PT, which are trained with different scales and architectures. We find that the benefits of neural scaling law are quite subtle in the stability and extensibility, in particular, their performance gaps are reduced marginally across different checkpoints. Instead, the baselines derived from vocabulary scaling laws obtain more significant benefits in stability and extensibility both on their high absolute accuracy and the reduced performance gaps compared with the original results.

6.2. Open-Vocabulary Fine-tuning

After verifying the vocabulary scaling laws, we turn to testify the superiority of SVFT under the open-vocabulary fine-tuning background. In the previous experiments, we find that fine-tuning baselines have already reached the high performances with the relatively unapparent gap in regular image recognition tasks, *e.g.*, ImageNet (Entity13), and ImageNet (Living17). Therefore, we resort to develop a more challenging setup to reflect the stability and the extensibility for CLIP-based fine-tuning models.

Benchmarks. Specifically, we provide the more challenging benchmarks **Birds** and **Rare Species** derived from the subsets in [25]. The Birds and Rare Species benchmarks were constructed to evaluate model performance under vocabulary expansion. Birds contains 525 fine-grained bird categories. Rare Species uses the unique scientific names of 400 rare species to avoid ambiguity. For the experiments, the augmenting vocabularies $\{V_i\}_{i=1}^{M}$ are denoted by a series of open-class set selected for training, each contains 25 classes for Birds or 20 classes for Rare Species, respectively. The remaining classes serve only as a pool of negative vocabulary items for expansion tests; their images are not used. For each training class in $\mathcal{V}^{(T)}$, 20 images were designated for training under a 16-shot learning setup, with the remaining images used for testing.

Baselines and evaluation. Distinct from Table.1 using pre-trained models to reflect the scaling law behaviors, the experiments on Birds and Rare Species also consider more challenging baselines for open-vocabulary fine-tuning: CoOp [36], CoCoOp [35], CLIP-Adapter [12], MAPLE [15]. They are evaluated along with the pre-trained CLIP and OpenCLIP fine-tuned with linear-probing to classify $\mathcal{V}^{(T)}$ and achieve the zero-shot inference on $U/\mathcal{V}^{(T)}$. Besides of our original model SVFT, we also consider its variant SVFT ($\mathcal{V}^{(T)}$) obtained by fine-tuning the class embedding on $\mathcal{V}^{(T)}$ (wo submodular class-name selection). To provide more comprehensive results on stability and extensiblity, all the baselines are evaluated with 10 target vocabulary scaling sequences drawn from the stochastic processes, then the stability and extensibilty with their deviation are plotted into the performance curves (refer to Fig.4).

Stability. On the Rare Species dataset, as the number of negative-class vocabularies grew from 20 to 400, SVFT kept accuracy within a 2-point drop, far outperforming all baselines. In contrast, the FSNL base model and other methods degraded substantially; even the best baseline, CLIP-Adapter, lost about 15 points. This sharp gap highlights SVFT's robustness to large-scale negative vocabularies. Results on the Birds dataset mirror this trend: across varying sizes of negative vocabularies, SVFT consistently sustains high accuracy. These experiments underscore two key strengths. First, SVFT offers strong vocabulary scalability, remaining effective as negative categories expand. Second, it delivers superior stability, preserving discriminative power and reliable classification under complex, open-vocabulary settings. Collectively, the evidence demonstrates SVFT's advancement in feature representation and class discrimination for open-world recognition.

Extensibility. Our proposed SVFT and SVFT ($\mathcal{V}^{(T)}$), demonstrate state-of-the-art extensibility. On both the Birds and the more challenging Rare Species benchmarks, SVFT consistently maintains the highest accuracy as the vocabulary of classes expands, proving its robust performance. This superior extensibility is a direct result of the model's exceptional stability. Its advanced ability to distinguish between classes and handle distractors allows it to manage a growing vocabulary more effectively than any baseline. While our models consistently lead, this advantage may appear less dramatic compared to the stability tests. This is because all models, including ours, inevitably experience performance degradation as the extensibility gradually including more open-class instances. The leading of SVFT is due to its decline mitigated more effectively than any competitor, yet the leading margin would be inevitably reduced

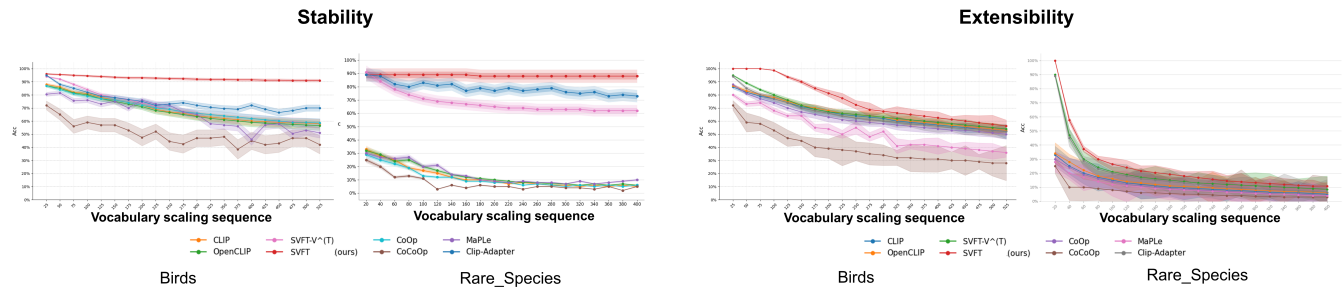

Figure 4. The curves of open-vocabulary fine-tuning baselines (*i.e.*, CLIP, OpenCLIP, CoOp, CoCoOp, CLIP-Adapter, MAPLE, and our SVFT-$\mathcal{V}^{(T)}$ (trained on $\mathcal{V}^{(T)}$ without data selection, refers to Class-name PT on $\mathcal{V}^{(T)}$) and SVFT) denote the variation on their stability and extensibility metrics derived from the challenging benchmarks, *i.e.*, Birds and Rare Species. We draw 10 target vocabulary sequence for each experiment, then evaluate the stability and extensibility by gradually increasing M.

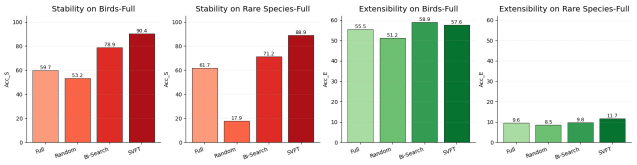

Figure 5. The ablation of SVFT using different algorithms to select class-name subset for each turn, and their stabilities and extensibilities change. SVFT indicates the linear greedy search in default.

with more open-class samples to join the test.

Full-class-name comparison. The fine-tuning information between SVFT and other open-vocabulary fine-tuning baselines are different: existing baselines are proposed to solve other problems such that $U/\mathcal{V}^{(T)}$ are supposed to be inaccessible whereas $U/\mathcal{V}^{(T)}$ solely denote a series of class names, which should have been accessible in practical scenarios. To this end, we provide the fair comparison between SVFT and the most competitive fine-tuning baselines, *i.e.*, MAPLE and CLIP-Adapter, with their fine-tuning includes full class names in $U/\mathcal{V}^{(T)}$ (see Appendix.B).We also conduct their evaluation with the vocabulary constructed by $\mathcal{V}^{(T)} \cup S^*$, where S^* indicates the class-name subset obtained by the submodular maximization (Eq.10) after the class-name embedding $\{e(w)\}$ have been well-trained in SVFT. The performances of MAPLE (35.00 in Birds, 86.75 in Rare Species) and CLIP-Adapter (94.76 in Birds, 91.07 in Rare Species) have been improved by incorporating the class names in $U/\mathcal{V}^{(T)}$ during fine-tuning, whereas their performances remain far behind SVFT (35.00 in Birds, 86.75 in Rare Species). Moreover, their performances on the "adversarial classes" selected by maximizing the training objective of SVFT, become disasters (MAPLE : 54.46 in Birds, 7.14 in Rare Species; CLIP-Adapter: 73.63 in Birds, 80.36 in Rare Species). It implies their fragility that SVFT does not suffer from (91.06 in Birds, 87.50 in Rare Species).

Class-name subset selection analysis. To analyze how the selection algorithms affect the performances, we compare the greedy search with the other three baselines, *i.e.*, Random (randomly select the same number of class names in U); Bi-Search [4]; Full (choose the full set of class names in the universe U). To ensure the fairness, we first run SVFT to obtain the optimal model with the selection records on the execution number and their training timelines, then perform the consistent optimization schedules by Random and Bi-Search. Fig.5 show that the linear greedy search is the optimal class-name selection algorithm to SVFT.

7. Limitation and Future Work

Several limitations should be acknowledged. Our framework assumes a pre-defined, finite vocabulary universe U, yet the space of possible class names is virtually unbounded in practice. If U fails to cover semantically confusing classes at deployment, the theoretical lower bound may become loose. Additionally, the greedy submodular selection requires evaluating marginal gains over all candidates in $U/\mathcal{V}^{(T)}$ at each step, creating a computational bottleneck for large-scale universes. Moreover, exclusively fine-tuning class-name embeddings while freezing encoders limits representational capacity for domains that diverge substantially from CLIP's pre-training distribution. Finally, the orthogonality regularization is a tractable surrogate but not a sufficient condition for extensibility, as near-orthogonality can be trivially satisfied in high-dimensional spaces among semantically related classes.

These limitations suggest promising future directions: leveraging large language models for dynamic vocabulary construction to relax the static universe assumption; adopting lazy or stochastic greedy strategies to reduce selection cost at scale; investigating joint class-name and lightweight vision-side adaptation while preserving the orthogonality constraint; incorporating taxonomic priors for more structured regularization; extending the vocabulary scaling law to open-vocabulary detection and segmentation; and integrating SVFT into continual learning loops for truly open-world vision–language systems.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 2

[2] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning*, pages 1298–1312. PMLR, 2022. 2

[3] Abhijit Bendale and Terrance Boult. Towards open world recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1893–1902, 2015. 2

[4] Niv Buchbinder, Moran Feldman, Joseph Seffi, and Roy Schwartz. A tight linear time (1/2)-approximation for unconstrained submodular maximization. *SIAM Journal on Computing*, 44(5):1384–1402, 2015. 5, 8

[5] Ziliang Chen, Pengxu Wei, Jingyu Zhuang, Guanbin Li, and Liang Lin. Deep cocktail networks: A universal framework for visual multi-source domain adaptation. *International Journal of Computer Vision*, 129(8):2328–2351, 2021. 2

[6] Ziliang Chen, Xin Huang, Quanlong Guan, Liang Lin, and Weiqi Luo. A retrospect to multi-prompt learning across vision and language. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22190–22201, 2023. 3

[7] Ziliang Chen, Yongsen Zheng, Zhao-Rong Lai, Quanlong Guan, and Liang Lin. Diagnosing and rectifying fake ood invariance: A restructured causal approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11471–11479, 2024. 2

[8] Ziliang Chen, Xin Huang, Xiaoxuan Fan, Keze Wang, Yuyu Zhou, Quanlong Guan, and Liang Lin. Reproducible vision-language models meet concepts out of pre-training. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14701–14711, 2025. 5, 6

[9] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023. 1, 2

[10] Xiuliang Duan, Dating Tan, Liangda Fang, Yuyu Zhou, Chaobo He, Ziliang Chen, Lusheng Wu, Guanliang Chen, Zhiguo Gong, Weiqi Luo, et al. Reason-and-execute prompting: Enhancing multi-modal large language models for solving geometry questions. In *Proceedings of the 32nd ACM international conference on multimedia*, pages 6959–6968, 2024. 2

[11] Satoru Fujishige. *Submodular functions and optimization*. Elsevier, 2005. 5

[12] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, pages 1–15, 2023. 3, 7

[13] Chuanxing Geng, Sheng-jun Huang, and Songcan Chen. Recent advances in open set recognition: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(10): 3614–3631, 2020. 2

[14] Lalit P Jain, Walter J Scheirer, and Terrance E Boult. Multi-class open set recognition using probability of inclusion. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part III 13*, pages 393–409. Springer, 2014. 2

[15] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19113–19122, 2023. 3, 7

[16] Andreas Krause and Daniel Golovin. Submodular function maximization. *Tractability*, 3(71-104):3, 2014. 5

[17] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 2

[18] Fengmao Lv, Changru Nie, Jianyang Zhang, Guowu Yang, Guosheng Lin, Xiao Wu, and Tianrui Li. Rethinking the effect of uninformative class name in prompt learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8345–8354, 2024. 3, 5

[19] Sarah Parisot, Yongxin Yang, and Steven McDonagh. Learning to name classes for vision and language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23477–23486, 2023. 3, 5

[20] Jitendra Parmar, Satyendra Chouhan, Vaskar Raychoudhury, and Santosh Rathore. Open-world machine learning: applications, challenges, and opportunities. *ACM Computing Surveys*, 55(10):1–37, 2023. 2

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2

[22] Shuhuai Ren, Lei Li, Xuancheng Ren, Guangxiang Zhao, and Xu Sun. Delving into the openness of clip. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9587–9606, 2023. 1, 2, 3, 6, 7, 11

[23] Walter J Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E Boult. Toward open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(7):1757–1772, 2012. 2

[24] Walter J Scheirer, Lalit P Jain, and Terrance E Boult. Probability models for open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 36(11):2317–2324, 2014. 2

[25] Samuel Stevens, Jiaman Wu, Matthew J Thompson, Elizabeth G Campolongo, Chan Hee Song, David Edward Carlyn, Li Dong, Wasila M Dahdul, Charles Stewart, Tanya Berger-Wolf, et al. Bioclip: A vision foundation model for the tree

of life. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19412–19424, 2024. 2, 7

[26] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 2

[27] Keze Wang, Liang Lin, Xiaopeng Yan, Ziliang Chen, Dongyu Zhang, and Lei Zhang. Cost-effective object detection: Active sample mining with switchable selection criteria. *IEEE transactions on neural networks and learning systems*, 30(3):834–850, 2018. 2

[28] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International conference on machine learning*, pages 23318–23340. PMLR, 2022. 2

[29] Ruijia Xu, Ziliang Chen, Wangmeng Zuo, Junjie Yan, and Liang Lin. Deep cocktail network: Multi-source unsupervised domain adaptation with category shift. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3964–3973, 2018. 2

[30] Xiaopeng Yan, Ziliang Chen, Anni Xu, Xiaoxi Wang, Xiaodan Liang, and Liang Lin. Meta r-cnn: Towards general solver for instance-level low-shot learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9577–9586, 2019. 2

[31] Tao Yu, Zhihe Lu, Xin Jin, Zhibo Chen, and Xinchao Wang. Task residual for tuning vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10899–10909, 2023. 3

[32] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14393–14402, 2021. 2

[33] Ji Zhang, Shihan Wu, Lianli Gao, Heng Tao Shen, and Jingkuan Song. Dept: Decoupled prompt tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12924–12933, 2024.

[34] Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021. 3

[35] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022. 1, 3, 7

[36] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 1, 3, 6, 7

[37] Jingyu Zhuang, Ziliang Chen, Junyi Zhang, Dongyu Zhang, and Zhaoquan Cai. Domain adaptation for retinal vessel segmentation using asymmetrical maximum classifier discrepancy. In *Proceedings of the ACM turing celebration conference-China*, pages 1–6, 2019. 2

[38] Jingyu Zhuang, Ziliang Chen, Pengxu Wei, Guanbin Li, and Liang Lin. Discovering implicit classes achieves open set domain adaptation. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 01–06. IEEE, 2022. 2

[39] Jingyu Zhuang, Ziliang Chen, Pengxu Wei, Guanbin Li, and Liang Lin. Open set domain adaptation by novel class discovery. *arXiv preprint arXiv:2203.03329*, 2022. 2

A. Appendix.A

In this section, we provide the proofs of Proposition.1 and Theorem.3.

A.1. The Equivalence between Equations 3,4 and their original presentations

Refer to [22], and we have the original presentations of stability and extensibility:

$$\begin{aligned}\text{Acc-S}\Big(\mathcal{V}^{(T)}, V^{(NT)}\Big) =& \mathbb{E}_{s\in\mathcal{S}_{M'}}\frac{1}{M}\sum_{i=1}^{M'}\text{Acc}\Big(\mathcal{V}^{(T)} \mid \mathcal{V}^{(T)}\cup\big(V_{s_1}^{(NT)}\cup\cdots\cup V_{s_i}^{(NT)}\big)\Big),\\ \text{Acc-E} =& \mathbb{E}_{s\in\mathcal{S}_{M'}}\frac{1}{M'}\sum_{i=1}^{M'}\text{Acc}\Big(V_{s_1}^{(T)}\cup\cdots\cup V_{s_i}^{(T)}\Big).\end{aligned} \tag{12}$$

Let compare them with Eq.3,4.

Stability. For $\text{Acc-S}\big(\mathcal{V}^{(T)}, V^{(NT)}\big)$ and $\text{ACC}_\text{S}(\mathcal{V}^{(T)})$, since s denotes the permutation set drawn from $\mathcal{S}_M$ to identify the negative vocabularies in $\{V_{s_1}^{(NT)},\cdots,V_{s_i}^{(NT)}\cdots,V_{s_M}^{(NT)}\}$, $\mathcal{S}_M$ can be equivalently considered as a set of negative vocabulary sequences, therefore it holds $\{V_j\}_{j=1}^{M}$ in Eq.3 refers to $\{\mathcal{V}^{(T)}, V_{s_1}^{(NT)},\cdots,V_{s_i}^{(NT)}\cdots V_{s_{M'}}^{(NT)}\}$ with respect to $\text{Acc-S}\big(\mathcal{V}^{(T)}, V^{(NT)}\big)$, when we set $M = M'+1$. We consider the $\mathcal{S}_{M'}$'s construction achieved by separating U into a set of sub-vocabularies $\{V_j\}_{j=1}^{i}\simeq\{\mathcal{V}^{(T)}, V_{s_1}^{(NT)},\cdots,V_{s_{i-1}}^{(NT)}\}$ then ranking them randomly (the randomness is denoted by our probability $\boldsymbol{P}_V$). So it leads to

$$\mathcal{V}_i^{(T)} = \cup_{j=1}^{i}V_j^{(T)} = \mathcal{V}^{(T)}\cup\big(V_{s_1}^{(NT)}\cup\cdots\cup V_{s_{i-1}}^{(NT)}\big) \tag{13}$$

then it is obvious that

$$\text{Acc}\Big(\mathcal{V}^{(T)} \mid V^{(T)}\cup\big(V_{s_1}^{(NT)}\cup\cdots\cup V_{s_i}^{(NT)}\big)\Big) = \mathbb{E}_{<\mathbf{x},\mathbf{y}>\sim\boldsymbol{D}_{\mathcal{V}^{(T)}}}\mathbb{I}\Big(\mathbf{y}=\underset{j\in\mathcal{V}_i^T}{\text{argmax}}\,P_{f,g}\big(\mathbf{x},\mathcal{V}_i^{(T)}\big)\Big). \tag{14}$$

So the stability in Eq.3 is equivalent with its original definition.

Extensibility. We can directly set $M = M'$ and $\forall j\in\{1,\cdots,M\}$, $V_j = V_{s_j}^{(T)}$. Therefore we have $\mathcal{V}_i^{(T)} = \cup_{j=1}^{i}V_j^{(T)} = V_{s_1}^{(T)}\cup\cdots\cup V_{s_i}^{(T)}$ then

$$\text{Acc}\Big(V_{s_1}^{(T)}\cup\cdots\cup V_{s_i}^{(T)}\Big) = \mathbb{E}_{<\mathbf{x},\mathbf{y}>\sim\boldsymbol{D}_{\mathcal{V}_i^{(T)}}}\mathbb{I}\Big(\mathbf{y}=\underset{j\in\mathcal{V}_i^{(T)}}{\text{argmax}}\,P_{f,g}\big(\mathbf{x},\mathcal{V}_i^{(T)}\big)\Big). \tag{15}$$

So the extensibility in Eq.3 is equivalent with its original definition.

A.2. Proof of Proposition.1

Proof. From the $\boldsymbol{P}_V$'s definition, we know $\mathcal{V}_i^T = \{V_j\}_{j=1}^{i}$ and $\mathcal{V}_{i+1}^T = \{V_j\}_{j=1}^{i+1} = \mathcal{V}_i^T\cup V_{i+1}, \forall i\in[M-1]$. Therefore

$$\begin{aligned}P_{f,g}^{(w_\mathbf{y})}(\mathbf{x},\mathcal{V}_{i+1}^{(T)}) =& \frac{\exp\left(\langle g(\boldsymbol{T}(e(w_\mathbf{y}))), f(\mathbf{x})\rangle/\gamma\right)}{\sum_{k=1}^{|\mathcal{V}_{i+1}^{(T)}|}\exp\left(\langle g(\boldsymbol{T}(e(w_k))), f(\mathbf{x})\rangle/\gamma\right)}\\ =& \frac{\exp\left(\langle g(\boldsymbol{T}(e(w_\mathbf{y}))), f(\mathbf{x})\rangle/\gamma\right)}{\sum_{w_k\in\mathcal{V}_{i+1}^{(T)}}\exp\left(\langle g(\boldsymbol{T}(e(w_k))), f(\mathbf{x})\rangle/\gamma\right)}\\ =& \frac{\exp\left(\langle g(\boldsymbol{T}(e(w_\mathbf{y}))), f(\mathbf{x})\rangle/\gamma\right)}{\sum_{w_k\in\mathcal{V}_i^{(T)}}\exp\left(\langle g(\boldsymbol{T}(e(w_k))), f(\mathbf{x})\rangle/\gamma\right)+\sum_{w_k\in V_{i+1}}\exp\left(\langle g(\boldsymbol{T}(e(w_k))), f(\mathbf{x})\rangle/\gamma\right)}\\ \leq& \frac{\exp\left(\langle g(\boldsymbol{T}(e(w_\mathbf{y}))), f(\mathbf{x})\rangle/\gamma\right)}{\sum_{w_k\in\mathcal{V}_i^{(T)}}\exp\left(\langle g(\boldsymbol{T}(e(w_k))), f(\mathbf{x})\rangle/\gamma\right)}\\ =& P_{f,g}^{(w_\mathbf{y})}(\mathbf{x},\mathcal{V}_i^{(T)}).\end{aligned} \tag{16}$$

The proposition has been proved. □

A.3. Proof of Theorem.3

Proof. Given embedding set $\{e(w)\}$ are fixed (denoted by $\{\overline{e}(w)\}$, consider $F(\{\overline{e}(w)\}, S)$

$$\begin{aligned} F(\{e(w)\}, S) :&= \mathbb{E}_{\substack{<\mathbf{x},\mathbf{y}> \\ \sim \boldsymbol{D}^{\mathsf{train}}_{\mathcal{V}^{(T)}}}} \Big[-\log P^{(w_{\mathbf{y}})}_{f,g}(\mathbf{x}, S \cup \mathcal{V}^{(T)}) + \lambda \mathbb{E}_{\substack{w \in \mathcal{V}^{(T)}, \\ w' \in S}} \Big(\langle g\big(\boldsymbol{T}(e(w))\big), g\big(\boldsymbol{T}(e(w'))\big) \rangle + 1 \Big) \Big] \\ &= \underbrace{\mathbb{E}_{\substack{<\mathbf{x},\mathbf{y}> \\ \sim \boldsymbol{D}^{\mathsf{train}}_{\mathcal{V}^{(T)}}}} -\log P^{(w_{\mathbf{y}})}_{f,g}(\mathbf{x}, S \cup \mathcal{V}^{(T)})}_{F_1(S)} + \underbrace{\lambda \mathbb{E}_{\substack{w \in \mathcal{V}^{(T)}, \\ w' \in S}} \langle g\big(\boldsymbol{T}(e(w))\big), g\big(\boldsymbol{T}(e(w'))\big) \rangle + \lambda |S||\mathcal{V}^{(T)}|}_{F_2(S)}. \end{aligned} \tag{17}$$

It is obvious that $F_2(S)$ is a modular function with respect to $\forall S \subset U/\mathcal{V}^{(T)}$, *i.e.*, $\forall S_1 \subset S_2, \forall w \in U/\{\mathcal{V}^{(T)} \cup S_2\}$,

$$\begin{aligned} F_2(S_1 \cup \{\hat{w}\}) - F_2(S_1) =& \lambda \mathbb{E}_{w' \in \mathcal{V}^{(T)}, \hat{w}} \langle g\big(\boldsymbol{T}(e(\hat{w}))\big), g\big(\boldsymbol{T}(e(w'))\big) \rangle + \lambda |\mathcal{V}^{(T)}| \\ =& F_2(S_2 \cup \{w\}) - F_2(S_2). \end{aligned} \tag{18}$$

It is also monotonic over the subset of $\mathcal{V}^{(T)}$ since $\forall S \subset \mathcal{V}^{(T)}, \forall w \in \mathcal{V}^{(T)}/S$, it holds

$$F_2(S \cup \{w\}) - F_2(S) = \lambda \mathbb{E}_{w' \in \mathcal{V}^{(T)}} \Big(\langle g\big(\boldsymbol{T}(e(w))\big), g\big(\boldsymbol{T}(e(w'))\big) \rangle + 1 \Big). \tag{19}$$

Observe that $\forall w' \in \mathcal{V}^{(T)}$, $\langle g\big(\boldsymbol{T}(e(w))\big), g\big(\boldsymbol{T}(e(w'))\big) \rangle + 1 \geq 0$. So $F_2(S \cup \{w\}) - F_2(S) \geq 0$ and $F_2(\cdot)$ is positively monotonic.

In terms of $F(\{e(w)\}, S) = F_1(S) + F_2(S)$, we only need to prove $F_1(S)$ is also positively monotonic and submodular. More specifically, we decompose $F_1(S)$ by the categories $w \in \mathcal{V}^{(T)}$, *i.e.*,

$$\begin{aligned} &F_1(S) = \sum_{w_{\mathbf{y}} \in \mathcal{V}^{(T)}} F_1(S; \mathbf{y}), \\ &\text{s.t. } F_1(S; \mathbf{y}) = -\mathbb{E}_{\mathbf{x} \sim \boldsymbol{D}^{\mathsf{train}}_{\mathcal{V}^{(T)}}(\cdot|\mathbf{y})} \frac{g(\boldsymbol{T}(e(w_{\mathbf{y}})))^\top f(\mathbf{x}))}{\gamma} + \left(\mathbb{E}_{\mathbf{x} \sim \boldsymbol{D}^{\mathsf{train}}_{\mathcal{V}^{(T)}}(\cdot|\mathbf{y})} \log \sum_{\forall w \in \hat{V}(w_{\mathbf{y}})} \exp\left(\frac{g(\boldsymbol{T}(e(w_{\mathbf{y}})))^\top f(\mathbf{x})}{\gamma} \right) \right). \end{aligned} \tag{20}$$

To this, we consider the set function $\log \sum_{i \in S} \exp(s_i)$ with the universe S. If $S_2 \subset S_1 \subset \mathsf{S}$ and $\forall s \notin S_1$, it holds

$$\begin{aligned} \log \sum_{i \in S_1 \cup \{s\}} \exp(s_i) - \log \sum_{i \in S_1} \exp(s_i) &= \log \left[\exp(s) + \sum_{i \in S_1} \exp(s_i) \right] - \log \left[\sum_{i \in S_1} \exp(s_i) \right] \\ &= \log \left[1 + \frac{\exp(s)}{\sum_{i \in S_1} \exp(s_i)} \right] \\ &= \log \left[1 + \frac{\exp(s)}{\sum_{i \in S_2} \exp(s_i) + \sum_{j \in S_1/S_2} \exp(s_j)} \right] \\ &\leq \log \left[1 + \frac{\exp(s)}{\sum_{i \in S_2} \exp(s_i)} \right] \\ &= \log \sum_{i \in S_2 \cup \{s\}} \exp(s_i) - \log \sum_{i \in S_2} \exp(s_i). \end{aligned} \tag{21}$$

From the derivation, we know that $\log \sum_{i \in S} \exp(s_i)$ is a submodular function with respect to the set S, in other words, $\log \sum_{\forall w \in \hat{V}(w_{\mathbf{y}})} \exp\left(\frac{g(\boldsymbol{T}(e(w_{\mathbf{y}})))^\top f(\mathbf{x})}{\gamma} \right)$ is submodular about the set of words with $U/\mathcal{V}^{(T)}$ as its universe. The derivation is satisfied for all $\mathbf{x} \sim \boldsymbol{D}^{\mathsf{train}}_{\mathcal{V}^{(T)}}(\cdot|\mathbf{y})$, with regards to the additive property, $\mathbb{E}_{\mathbf{x} \sim \boldsymbol{D}^{\mathsf{train}}_{\mathcal{V}^{(T)}}(\cdot|\mathbf{y})} \log \sum_{\forall w \in \hat{V}(w_{\mathbf{y}})} \exp\left(\frac{g(\boldsymbol{T}(e(w_{\mathbf{y}})))^\top f(\mathbf{x})}{\gamma} \right)$ is also submodular with $U/\mathcal{V}^{(T)}$ as its universe. Note that $-\mathbb{E}_{\mathbf{x} \sim \boldsymbol{D}^{\mathsf{train}}_{\mathcal{V}^{(T)}}(\cdot|\mathbf{y})} \frac{g(\boldsymbol{T}(e(w_{\mathbf{y}})))^\top f(\mathbf{x}))}{\gamma}$ is constant with respect to the set variable define in the subset of $U/\mathcal{V}^{(T)}$. So $F_1(S; \mathbf{y})$

is a submodular function, and their linear combination over the categories in $\mathcal{V}^{(T)}$, *i.e.*, $F_1(S;\mathbf{y}) = \sum_{y\in\mathcal{V}^{(T)}} F_1(S;\mathbf{y})$, is also submodular.

Beyond this, $\forall S \subset \mathcal{V}^{(T)}$, $\forall w \in \mathcal{V}^{(T)}/S$, it holds

$$\begin{aligned} F_1(S\cup\{w\}) - F_1(S) =& \sum_{w_{\mathbf{y}}\in\mathcal{V}^{(T)}} F_1(S\cup\{w\};\mathbf{y}) - F_1(S;\mathbf{y}) \\ =& \mathbb{E}_{w_{\mathbf{y}}\in\mathcal{V}^{(T)},\mathbf{x}\sim\boldsymbol{D}^{\text{train}}_{\mathcal{V}^{(T)}}(\cdot|\mathbf{y})} \left(\log\left[1 + \frac{\exp\left(\frac{g(\boldsymbol{T}(e(w)))^\top f(\mathbf{x})}{\gamma}\right)}{\sum_{\forall\hat{w}\in\hat{V}(w_{\mathbf{y}})}\exp\left(\frac{g(\boldsymbol{T}(e(\hat{w})))^\top f(\mathbf{x})}{\gamma}\right)} \right] \right) > 0, \end{aligned} \tag{22}$$

so we know $F_2(\cdot)$ is also positively monotonic.

Therefore, $F(\{e(w)\},S) = F_1(S) + F_2(S)$ is submodular and positively monotonic with respect to the universe $U/\mathcal{V}^{(T)}$, and $\max_{S\subset U/\mathcal{V}^{(T)},|S|\leq K} F(\{\overline{e}(w)\},S)$ a submodular maximization problem with the K-size cardinality constraint, and its global optimal subset S^* can be approximated with $\hat{S}$ obtained by a greedy-search algorithm:

$$F(\{\overline{e}(w)\},\hat{S}) \geq (1-\frac{1}{e})F(\{\overline{e}(w)\},S^*). \tag{23}$$

□

B. Appendix.B

B.1. Implementation

In our implementation, we separately achieve the submodular subset selection with classes in $\mathcal{V}^{(T)}$, thus,

$$\max_{S_{\mathbf{y}}\subset U/\mathcal{V}^{(T)},|S_{\mathbf{y}}|\leq K} F(\{e(w)\},S_{\mathbf{y}}) = F_1(S_{\mathbf{y}};\mathbf{y}) + \lambda\mathbb{E}_{w_{\mathbf{y}},w'\in S}\langle g\big(\boldsymbol{T}(e(w_{\mathbf{y}}))\big), g\big(\boldsymbol{T}(e(w'))\big)\rangle$$

which is obviously also submodular. To this, it refers to $\mathbf{m}$ class-name set $\{S_{\mathbf{y}}\}_{w_{\mathbf{y}}\in\mathcal{V}^{(T)}}$ where each denotes the submodular subset drawn with the formula above. For this, we refine the regular cross-entropy loss into

$$\min_{\{e(w)\}_{\forall w\in\mathcal{V}^{(T)}}} \frac{1}{|\mathcal{V}^{(T)}|} \sum_{w_{\mathbf{y}}\in\mathcal{V}^{(T)}} \min_{\{e(w)\}_{\forall w\in S_{\mathbf{y}}}} \mathbb{E}_{\substack{<\mathbf{x},\mathbf{y}> \\ \sim\boldsymbol{D}^{\text{train}}_{\mathcal{V}^{(T)}}}} -\log P^{(w_{\mathbf{y}})}_{f,g}(\mathbf{x}, S_{\mathbf{y}}\cup\mathcal{V}^{(T)}) \tag{24}$$

which aims to optimize the class-name embedding both in $\mathcal{V}^{(T)}$ and the selected open classes.

In all experiments, we setup $K = |\mathcal{V}^{(T)}| - 1$ and the class-name subset selection is executed at the beginning of each epoch; we setup $\lambda = 0$ if we execute the stability experiments yet setup $\lambda = 1$ if we execute the extensibility experiments in this paper. We use a single Nvidia A100 GPU to finish all experiments, and the wall-clock time is 2 hours for Rare Species and 5 hours for Birds.

The algorithm is shown in Algo.1

B.2. Experimental Setup

Here we provide the experimental setups in our first and second experiments.

Backbone Models, Benchmarks, and Baselines: The backbone model for our first and the second experiments is CLIP and OpenCLIP (ViT-B/16), pre-trained on the LAION-400M dataset. In the first experiment, we adopt the benchmarks and baselines consistent with . The second experiment is conduct on the following two benchmarks:

- *Birds*: Contains 525 distinct bird species. We randomly select 25 classes for training and use the remaining 500 classes as distractor (negative) vocabulary for evaluation.
- *Rare Species*: Contains 400 rare species. We randomly select 20 classes for training, with the remaining 380 classes serving as the distractor vocabulary. For both datasets, we follow a 16-shot learning protocol. For each training class, 16 image samples are used for fine-tuning, and the rest are reserved for the test set. The images corresponding to the distractor vocabulary are not used in either training or testing.

For the Birds dataset, we add 25 distractor classes at each step, up to the full set of 525 classes. For the Rare Species dataset, we add 20 distractor classes at each step, up to the full set of 400 classes. This protocol allows us to measure the stability and robustness of each model as the size of the negative vocabulary increases, simulating a real-world open-vocabulary scenario.

Algorithm 1: Example algorithm

Input: Pre-trained CLIP encoders f,g; $\mathcal{V}^{(T)}$, $U/\mathcal{V}^{(T)}$, $\boldsymbol{P}_V$
Parameter: class-name embedding $\{e(w)\}$
Output: Fine-tuned class-name embedding $\{e^*(w)\}$

while per epoch **do**
  $\forall w_{\mathbf{y}} \in \mathcal{V}^{(T)}$, $\hat{S}_{\mathbf{y}} = \arg\max_{S_{\mathbf{y}} \subset U/\mathcal{V}^{(T)}, |S_{\mathbf{y}}| \leq K} F(\{e(w)\}, S_{\mathbf{y}})$ (Parallel executed across classes in $\mathcal{V}^{(T)}$).
  while $B \sim \boldsymbol{D}^{\mathsf{train}}_{\mathcal{V}^{(T)}}$ **do**
    $\min_{\{e(w)\}_{\forall w \in \mathcal{V}^{(T)}}} \frac{1}{|\mathcal{V}^{(T)}|} \sum_{w_{\mathbf{y}} \in \mathcal{V}^{(T)}} \min_{\{e(w)\}_{\forall w \in S_{\mathbf{y}}}} \mathbb{E}_{\substack{<\mathbf{x},\mathbf{y}> \\ \sim \boldsymbol{D}^{\mathsf{train}}_{\mathcal{V}^{(T)}}}} - \log P^{(w_{\mathbf{y}})}_{f,g}(\mathbf{x}, S_{\mathbf{y}} \cup \mathcal{V}^{(T)})$
  end while
  Update $\{e(w)\}_{\forall w \in \mathcal{V}^{(T)} \cup \{S_{\mathbf{y}}\}_{w(\mathbf{y})}}$.
end while
return $\{e^*(w)\}$

Then we compare our SVFT method against several baselines, including the zero-shot performance of CLIP and OpenCLIP, as well as state-of-the-art parameter-efficient fine-tuning (PEFT) methods: FSNL (our base model), CoOp, CoCoOp, CLIP-Adapter, and MaPle. Note that MaPle is implemented on the original CLIP (ViT-B/16) backbone due to its architectural modifications to both the vision and language encoders.

Hard-to-Classify Set Selection: Before fine-tuning, we construct a hard-to-classify vocabulary set for each of the $|\mathcal{V}^{(T)}|$ training classes. This is modeled as a submodular maximization problem, which we solve using a greedy algorithm. The objective function for the greedy selection is the FSNL model's loss, which effectively identifies the most confusing vocabulary for each target class. This stage is computationally intensive and is performed as a pre-processing step.

Parallel Fine-Tuning: We fine-tune the model using the 16-shot training data.

Optimizer: We use Stochastic Gradient Descent (SGD). Learning Rate: The base learning rate is set to 2e-4 for a batch size of 128 and is scaled linearly with the actual batch size. A cosine annealing schedule is used for learning rate decay.

Batch Size: To enable efficient parallel processing of hard-set losses, the batch size is set to $|\mathcal{V}^{(T)}|n$ shots. For example, for the Birds dataset, the batch size is 25 × 16 = 400.

Epochs: The model is fine-tuned for 200 epochs.

Evaluation Protocol: To assess performance under vocabulary expansion, we measure top-1 classification accuracy on the test set. The evaluation begins with a vocabulary consisting only of the $|\mathcal{V}^{(T)}|$ training classes. We then incrementally add distractor classes from the negative vocabulary pool.